REDUCED BASIS ERROR BOUND COMPUTATION OF PARAMETER-DEPENDENT NAVIER–STOKES EQUATIONS BY THE NATURAL NORM APPROACH∗

SIMONE DEPARIS†

Abstract. This work focuses on the a posteriori error estimation for the reduced basis method applied to partial differential equations with quadratic nonlinearity and affine parameter dependence. We rely on natural norms—local parameter-dependent norms—to provide a sharp and computable lower bound of the inf-sup constant. We prove a formulation of the Brezzi–Rappaz–Raviart existence and uniqueness theorem in the presence of two distinct norms. This allows us to relax the existence condition and to sharpen the field variable error bound. We also provide a robust algorithm to compute the Sobolev embedding constants involved in the error bound and in the inf-sup lower bound computation. We apply our method to a steady natural convection problem in a closed cavity, with a Grashof number varying from 10 to 107.

Key words. reduced basis methods, a posteriori error estimation, Brezzi–Rappaz–Raviart theory, steady incompressible Navier–Stokes equations, natural convection

AMS subject classifications. 35Q30, 46E35, 65N30, 76D03, 76D05

DOI. 10.1137/060674181

1. Introduction. In this work we are interested in the numerical approximation of parameter- (μ-) dependent quadratic nonlinear partial differential equations—in particular the natural convection problem for Navier–Stokes equations—and the prediction of an “output of interest” which is a functional of the field variable ue(μ),

(1.1) sμ(μ) = ℓ(ue(μ)) ∈ ℝ,

where ℓ is a continuous linear form. The solution ue(μ) may not be unique; nevertheless, we consider only one solution branch; hence we presume local uniqueness.

The outputs are related to energies or forces, stresses or strains, flow rates or pressure drops, temperatures or fluxes and are functions of an “input” parameter P-vector μ ∈ D ⊂ ℝP, which is related to geometry, physical properties, boundary conditions, or loads.

We directly adopt the description and notation of Sen et al. [32], where the natural norms are introduced and applied to the error bound computation of coercive and noncoercive linear elliptic partial differential equations.

The field variable ue(μ) ∈ Xc—say, velocity, pressure, or temperature—satisfies in weak form the μ-parametrized nonlinear partial differential equation

(1.2) a(ue(μ), v; μ) = f(v) ∀ v ∈ Xc.

Here Xc is the appropriate function space with norm || · ||x, a(·, ·; μ) is continuous, quadratic in the first variable, and linear in the second one, and f is a continuous linear form.

∗Received by the editors November 7, 2006; accepted for publication (in revised form) December 12, 2007; published electronically April 23, 2008. This work was supported by DARPA and AFOSR under grant FA9550-05-1-0114, by the Singapore-MIT Alliance, and by PSC grant ASC040010P. http://www.siam.org/journals/sinum/46-4/67418.html
†Mechanical Engineering Department, Massachusetts Institute of Technology, Cambridge, MA 02139 (simone.deparis@epfl.ch).
Unfortunately, to achieve the desired accuracy, the evaluation $\mu \rightarrow s^e(\mu)$ by discrete projection methods such as finite element or spectral methods is simply too costly in the many-query and real-time contexts often of interest in engineering. Low-order models—we consider here reduced basis approximations—are increasingly popular in the engineering analysis, parameter estimation, design optimization, and control contexts.

In the reduced basis approach [1, 18, 6, 22, 20, 13, 23, 11, 12, 25], we approximate $s^e(\mu), u^e(\mu)$ with $s_N(\mu), u_N(\mu)$: Given $\mu \in \mathcal{D}$,

$$s_N(\mu) = \ell(u_N(\mu)),$$

where $u_N(\mu) \in W_N$ satisfies

$$a(u_N(\mu), v; \mu) = f(v) \quad \forall v \in W_N.$$

Here $W_N$ is a problem-specific space of dimension $N \leq N_{\text{max}}$ that focuses on the (typically very smooth) parametric manifold of interest—${u^e(\mu) \mid \mu \in \mathcal{D}}$—and thus the reduced basis method enjoys very rapid convergence as $N$ increases [6, 15]; $N_{\text{max}}$ is the maximal size of our reduced basis.

Our own effort is dedicated to the development of a posteriori error estimators for reduced basis approximations in the nonlinear case [13, 23, 35, 17, 11, 12, 24]. Since, in general, we cannot find the exact (our superscript “e” above) solution, we replace $s^e(\mu), u^e(\mu)$ with a Galerkin finite element approximation

$$s_N(\mu), u_N(\mu).$$

$N$ represents the dimension of our finite element space.

For some fixed sufficiently large $N = N_t$, we presume that the finite element and the exact solution are very close. We shall call $u_{N_t}(\mu)$ a “truth” solution, upon which we provide inexpensive—complexity independent of $N_t$—and sharp error bounds $\Delta_N(\mu)$ and $\Delta^s_N(\mu)$ such that

$$\|u_{N_t}(\mu) - u_N(\mu)\|_X \leq \Delta_N(\mu) \quad \text{and}$$

$$|s_{N_t}(\mu) - s_N(\mu)| \leq \Delta^s_N(\mu) \quad \forall \mu \in \mathcal{D}.$$

In the absence of such rigorous error bounds we cannot efficiently determine if $N$ is too small—and our reduced basis approximation unacceptably inaccurate—or if $N$ is too large—and our reduced basis approximation unnecessarily expensive—or we cannot establish the very existence of a “truth” solution $u_{N_t}(\mu)$ [8, 34, 33]. We cannot determine in “real time” if critical design conditions and constraints are satisfied. In fact, we cannot even construct an efficient and well-conditioned reduced basis approximation space $W_N$ [17, 16].

The dramatic dimension reduction—from $N_t$ to $N$—in conjunction with offline/online computational procedures [13, 23, 2, 9], yields very large savings in the many-query and real-time contexts: The online complexity depends only on the size of the reduced basis space $N$, which is typically orders of magnitude smaller than the dimension of the finite element space $N_t$. Hence we may consider a highly accurate “truth” approximation, at the detriment of the offline complexity only.

Relative to earlier work [35, 17, 34, 33], we adapt the natural norms technique proposed in [32] to the quadratic case in order to improve the construction and evaluation of a lower bound of the inf-sup constant (which we shall subsequently call the

---

1 For simplicity, we consider a purely primal approach; we shall subsequently describe also a primal-dual formulation.

2 Other projection methods, such as spectral or spectral elements methods, may be used as well; cf. [11, 12].
“inf-sup lower bound”). We also extend the Brezzi–Rappaz–Raviart theory for the case of mixed norms (in our case, $X$ and natural norms); this enables one to better control the effectivity of our error bound and to significantly reduce the constraint on the dual norm of the residual for the existence and uniqueness of a (finite element) solution in the neighborhood of our reduced basis approximation.

The effectivity relative to the natural norm of the field variable error is proven to be very accurate: locally of order one when the dual norm of the residual is small. However, when considering only a primal formulation—as in the example that we provide—we have no control of the effectivity on the output. To partially address this issue, we also propose a primal-dual formulation, whose apparent advantages would require further investigation, in particular in conjunction with deflation.

Our improvements enable a feasible and fast inf-sup lower bound construction based on a relatively small number of local norms. The inf-sup lower bound is of order one in the local natural norm and hence is intrinsically much simpler to compute than in existing works [34, 17, 33]. The lower bound components are also simplified: They rely on the computation of Sobolev embedding constants and eigenvalues with respect to the natural norms. We develop robust tools to compute them based on multiple initial guesses modes and, regarding the Sobolev embedding constants, the restriction to a small discrete subspace to provide the initial guess and on a fixed point algorithm.

As a result, we are now able to cover much larger parameter sets than in earlier works. For example, in [17, 33] the authors considered natural convection with a Grashof number (ratio buoyancy over viscosity) from 10 to $10^4$. Instead, we have applied the reduced basis method for a Grashof number up to $10^7$ without increasing—actually by simplifying—the complexity of the error bound computation. To enable the finite element resolution of our problem at these values of the Grashof number and to speed up the offline work, we implemented a parallel code and a domain decomposition preconditioner based on a sequential one proposed in [5].

We first introduce the basic assumptions that we are going to use throughout the paper and define our natural norms for nonlinear partial differential equations (section 2). We then prove an existence and unicity result based on the Brezzi–Rappaz–Raviart theory and provide an output error bound based on a pure primal approach (section 3). In section 4 we develop our inf-sup lower bound construction that enables a feasible offline/online computational strategy (section 5). We then apply our approach to a steady natural convection problem (section 6), provide the numerical results (section 7), and present our conclusions (section 8). The appendix is devoted to the output error bound by a primal-dual formulation.

2. Problem statement.

2.1. Reduced basis formulation. For any $\mu \equiv (\mu_1, \ldots, \mu_P)$ in the closed input domain $D \subset \mathbb{R}^P$, $a(\cdot, \cdot; \mu) : X^e \times X^e \to \mathbb{R}$ is a parameter-dependent form, which is quadratic with respect to its first argument and linear with respect to the second one. The forms $\ell$ and $f : X^e \to \mathbb{R}$ are parameter-independent and linear (extension to parameter-dependent linear forms is straightforward; we provide also an extension to quadratic outputs).

Our “truth” or “reference” finite element approximation to the exact output and field variable $s(\mu) \equiv s^{N}(\mu)$ and $u(\mu) \equiv u^{N}(\mu) \in X^{N} \equiv X$, for a given $\mu \in D$,

$$
s(\mu) = \ell(u(\mu)) \quad \text{and} \quad a(u(\mu), v; \mu) = f(v) \quad \forall \ v \in X.
$$ (2.1)
We assume that \( \mathcal{N}_i \) is chosen sufficiently large so that \( s(\mu) \) and \( u(\mu) \) are essentially indistinguishable from \( s^*(\mu) \) and \( u^*(\mu) \), respectively. We shall build our reduced basis approximation upon this “truth” approximation, and we shall evaluate the error in our reduced basis approximation with respect to this “truth” approximation. The online complexity (and stability) of our reduced basis approach is independent of \( \mathcal{N}_i \) (cf. [17, 33, 32]); hence, we may choose \( \mathcal{N}_i \) to be “arbitrarily” large at no detriment to (online) performance.

We shall suppose that our form \( a \) is “affine” in the parameter: For some fixed integers \( Q_0 \) and \( Q_1 \)—typically larger than \( P \), sometimes by a considerable factor—we require that

\[
(2.2) \quad a(u, v; \mu) = \sum_{q=1}^{Q_0} \Theta^q_0(\mu) a^q_0(u, v) + \frac{1}{2} \sum_{q=1}^{Q_1} \Theta^q_1(\mu) a^q_1(u, v) \quad \forall \, u, v \in X, \forall \, \mu \in D,
\]

where \( \Theta^q_0, \Theta^q_1 : D \to \mathbb{R}, a^q_0 : X \times X \to \mathbb{R}, 1 \leq q \leq Q_0 \), and \( a^q_1 : X \times X \times X \to \mathbb{R}, 1 \leq q \leq Q_1 \), are parameter-dependent functions and parameter-independent continuous bilinear and trilinear forms, respectively. (In particular, this assumption presumes a quadratic nonlinearity in our partial differential equation.) Without loss of generality, we assume that, for \( 1 \leq q \leq Q_1 \) and \( \forall \, u, v, w \in X \), \( a^q_1(u, v, w) = a^q_1(w, u, v) \). We shall further assume that \( \Theta^q_i \in C^1(D) \), \( 1 \leq q \leq Q_i, i = 0, 1 \).

We denote the inner product and norm associated with our Hilbert space \( X = X^{N_i} \), finite-dimensional) as \( (w, v)_X \) and \( \|v\|_X = \sqrt{(v, v)_X} \), respectively. We further define the dual norm for any bounded linear functional \( f \) as

\[
\|f\|_{X'} = \sup_{w \in X} \frac{|f(v)|}{\|v\|_X}.
\]

For example, when considering two-dimensional Navier–Stokes equations, the exact space \( X^e \) shall satisfy \( H^0_0(\Omega) \times H^1_0(\Omega) \times L^2_0(\Omega) \subset X^e \subset H^1(\Omega) \times H^1(\Omega) \times L^2(\Omega) \): Here \( \Omega \subset \mathbb{R}^2 \) is a spatial domain with suitably regular boundary \( \partial \Omega \); \( L^2(\Omega) \) is the Hilbert space of integrable functions; \( H^1(\Omega) \) is the usual Hilbert space of derivative-square-integrable functions; \( H^0_0(\Omega) \) indicates the functions with null trace on \( \partial \Omega \) and \( L^2_0(\Omega) \) the zero mean ones. The typical choice for our inner product \( (\cdot, \cdot)_X \) is the standard \( H^1(\Omega) \times H^1(\Omega) \times L^2(\Omega) \) inner product.

If in this example the Reynolds number is the parameter, then the affine decomposition (2.2) becomes, for \( \mu = (\text{Re}), u = (u_1, u_2, p), v = (v_1, v_2, q), Q_0 = 2, \) and \( Q_1 = 1 \),

\[
a^1_0(u, v) = -\int_{\Omega} p \partial_i v_i - \int_{\Omega} q \partial_i u_i + \lambda \int_{\Omega} p + \gamma \int_{\Omega} q, \quad \Theta^1_0(\mu) = 1,
\]

\[
a^0_1(u, v) = \int_{\Omega} \partial_i u_j \partial_i v_j, \quad \Theta^0_1(\mu) = \text{Re}^{-1},
\]

\[
a^1_1(u, z, v) = \int_{\Omega} u_j \partial_i z_i v_j + \int_{\Omega} z_j \partial_i u_i v_j, \quad \Theta^1_1(\mu) = 1
\]

(with summation over repeated indices and partial derivatives with respect to \( x_1, x_2 \)).

By recalling (2.2) and the symmetry of the \( a^q_i \)'s in the first two variables, the Fréchet derivative of \( a(\cdot, \cdot; \mu) \) with respect to the first variable at a point \( w \in X \) can be expressed as

\[
da(w; \mu)(u, v) = \sum_{q=1}^{Q_0} \Theta^q_0(\mu) a^q_0(u, v) + \sum_{q=1}^{Q_1} \Theta^q_1(\mu) a^q_1(u, w, v) \quad \forall \, u, v \in X, \forall \, \mu \in D.
\]
For any $\mu$ in $D$ and any solution $u(\mu)$ of (2.1) in the region of our interest, we assume that $da(u(\mu)\mu)$ is “stable” and continuous in the sense that there exist $\beta_0 > 0$ and $\gamma_0 \in \mathbb{R}$ such that, $\forall \mu \in D$,

\begin{align}
0 < \beta_0 < \beta(\mu) &\equiv \inf_{w \in X} \sup_{v \in X} \frac{da(u(\mu)\mu)(w, v)}{\|w\|_X \|v\|_X}, \\
\infty > \gamma_0 > \gamma(\mu) &\equiv \sup_{w \in X} \sup_{v \in X} \frac{da(u(\mu)\mu)(w, v)}{\|w\|_X \|v\|_X}.
\end{align}

(2.3)

It then follows that in the neighborhood of $u(\mu)$ the solution to (2.1) is unique. We further assume that $\ell$ and $f$ are in $X'$—bounded linear functionals.

### 2.2. Natural norms. Let $u_N(\mu)$ be the reduced basis approximation to $u(\mu)$ given by (1.3). We next introduce [35, 17, 7, 14, 32] the parametrized linear operator $T_N^\mu : X \rightarrow X$ such that

\begin{equation}
(T_N^\mu w, v)_X = da(u_N(\mu)\mu)(w, v), \quad \forall w, v \in X.
\end{equation}

(2.4)

Our method—in particular our inf-sup lower bound construction—requires a discrete set of $K$ parameter values $V^K = \{\mathbf{\mu}_1, \ldots, \mathbf{\mu}_K\} \subset D$—upon which to construct local norms—a fixed integer $N \leq N_{\max}$, and an indicator function $I^K : D \rightarrow V^K$ which associates to any $\mu$ in $D$ a member of $V^K$. Typically, $I^K$ defines the “closest”—in a sense that we have to define—element of $D$.

We assume that, $\forall \mathbf{\mu} \in V^K$,

\begin{align}
0 < \beta_I(\mathbf{\mu}) &\equiv \inf_{w \in X} \sup_{v \in X} \frac{da(u_\mathbf{\mu}(\mathbf{\mu})\mathbf{\mu})(w, v)}{\|w\|_X \|v\|_X} \equiv \inf_{w \in X} \frac{\|T_{\mathbf{\mu}}^\mathbf{\mu}w\|_X}{\|w\|_X}, \\
\infty > \gamma_I(\mathbf{\mu}) &\equiv \sup_{w \in X} \sup_{v \in X} \frac{da(u_\mathbf{\mu}(\mathbf{\mu})\mathbf{\mu})(w, v)}{\|w\|_X \|v\|_X} \equiv \sup_{w \in X} \frac{\|T_{\mathbf{\mu}}^\mathbf{\mu}w\|_X}{\|w\|_X}.
\end{align}

(2.5)

Note that this is true if $u_\mathbf{\mu}(\mathbf{\mu})$ is close enough to $u(\mathbf{\mu})$ (even though this is not required here); from the Cauchy–Schwarz inequality, $v = T_{\mathbf{\mu}}^\mathbf{\mu}w$ is the inner supremizer in (2.5) and (2.6).

We then introduce, for any given $\mathbf{\mu} \in V^K$ and $\mathbf{\mu}$, our “natural norm”

\begin{equation}
\|v\|_\mathbf{\mu}^\mathbf{\mu} \equiv \|T_{\mathbf{\mu}}^\mathbf{\mu}v\|_X, \quad \forall v \in X.
\end{equation}

(2.7)

(To simplify the notation we have dropped the index $\mathbf{\mu}$ from the norm symbol.) This norm is the extension of the natural norm introduced in [32] for the linear case. Note that, thanks to our assumptions on $\beta_I(\mathbf{\mu})$ and $\gamma_I(\mathbf{\mu})$, (2.7) does indeed define a norm, which is equivalent to $\| \cdot \|_X$.

### 2.3. Trilinear forms continuity constants. In the development of our error bound, we assume that there exists a positive $\rho_I(\mathbf{\mu})$ such that

\begin{equation}
|da(z^2; \mu)(v, w) - da(z^1; \mu)(v, w)| \leq \rho_I(\mathbf{\mu}) \|z^2 - z^1\|_\mathbf{\mu} \|v\|_\mathbf{\mu} \|w\|_X
\end{equation}

for all $z^1$, $z^2$, $v$, and $w$ in $X$.

---

3In practice, $\mathbf{\mu}$ may depend on $\mathbf{\mu}$, and the results of this work still hold.

4The process by which we select “good” $V^K$ and $I^K$ is described in section 5.
Similarly, in the construction of an inf-sup lower bound, we assume that there exists a positive $\rho_{X,\|\cdot\|}$ such that

$$
(2.9) \quad \left| \sum_{q=1}^{Q_1} \Theta_q^2(\rho) a_q^2(v, z, w) \right| \leq \rho_{X,\|\cdot\|} \|z\|_X \|v\|_\rho \|w\|_X,
$$

$$
(2.10) \quad \left| \sum_{q=1}^{Q_1} (\Theta_q^2(\mu) - \Theta_q^2(\rho)) a_q^2(v, z, w) \right| \leq \max_{q=1,...,Q_1} \frac{|\Theta_q^2(\mu) - \Theta_q^2(\rho)|}{|\Theta_q^2(\rho)|} \rho_{X,\|\cdot\|} \|z\|_X \|v\|_\rho \|w\|_X
$$

for all $z, v,$ and $w$ in $X$.

To a given problem, we have to provide these constants; in sections 6.3 and 6.5 we show how to compute these constants, which are related to the Sobolev embedding $(3.1)$ and $(3.3)$.

To prove the uniqueness result in the neighborhood—with respect to the natural norms—of a given reduced basis approximation. Let $u \in X$, such that $u(\mu) = 0$. A fixed $\epsilon > 0$ is also that we explicitly know either $\beta_{\rho}(\mu)$ or a rigorous positive lower bound to it. The results in this section extend straightforwardly in the latter case by replacing $\beta_{\rho}(\mu)$ by its lower bound.

### 3. Existence and uniqueness.

**Lemma 3.1.** Assume that $\beta_{\rho}(\mu) > 0$. There can be at most one solution $u(\mu)$ to (2.1) such that

$$
\left| u(\mu) \right| \in B_{\|\cdot\|} \left( u_N(\mu), \frac{\beta_{\rho}(\mu)}{\beta_{\rho}(\mu)} \right).
$$

**Proof.** The proofs of this lemma and the following one are based on the original proofs in [4] and are extended to the case of two distinct norms. Let $B_{\|\cdot\|}$ be an open ball around $u \in X$, of radius $\epsilon > 0$, and metric induced by $\|\cdot\|_{\rho}$.

$$\langle A(u; \mu), v \rangle = a(u, v; \mu) - f(v) \quad \forall u, v \in X,$$

$$\langle DA(z; \mu) u, v \rangle = da(z; \mu)(u, v) \quad \forall u, v \in X,$$

$$H(w; \mu) = w - DA(u_N(\mu); \mu)^{-1} A(w; \mu).$$

$DA(u_N(\mu); \mu)$ is invertible thanks to our assumption that $\beta_{\rho}(\mu) > 0$. A fixed point of $H$ must be a solution to (2.1) and vice versa; we are going to show that there exists at most one.
We consider $H(z^2, \mu) - H(z^1, \mu)$, which from the definition (3.4) can be expressed as

\begin{equation}
H(z^2, \mu) - H(z^1, \mu) = (z^2 - z^1) - DA(u_N(\mu); \mu)^{-1} (A(z^2; \mu) - A(z^1; \mu)).
\end{equation}

We observe that, since our nonlinearity is quadratic,

\begin{equation}
A(z^2; \mu) - A(z^1; \mu) = DA\left(\frac{z^1 + z^2}{2}; \mu\right)(z^2 - z^1).
\end{equation}

Applying (3.5) and (3.6) gives

\[
DA(u_N(\mu); \mu) \left( H(z^2; \mu) - H(z^1; \mu) \right)
= DA(u_N(\mu); \mu) (z^2 - z^1) - (A(z^2; \mu) - A(z^1; \mu))
= DA(u_N(\mu); \mu) - DA\left(\frac{z^1 + z^2}{2}; \mu\right) (z^2 - z^1).
\]

We then have, by recalling (2.8),

\[
\langle DA(u_N(\mu); \mu)(H(z^2; \mu) - H(z^1; \mu)), v \rangle
\leq \rho_{\overline{\Pi}}(\mu) \left\| u_N(\mu) - \frac{1}{2}(z^2 + z^1) \right\| \left\| z^2 - z^1 \right\|_{\overline{\Pi}} \| v \|_X.
\]

Furthermore, from the definition of $\beta_{\overline{\Pi}}(\mu)$, $T_N^\mu$, and $DA(u_N(\mu); \mu)$, ((3.1), (2.4), and (3.3), respectively), we have

\begin{equation}
\beta_{\overline{\Pi}}(\mu) \left\| H(z^2; \mu) - H(z^1; \mu) \right\|_{\overline{\Pi}} T_N^\mu (H(z^2; \mu) - H(z^1; \mu))_X
\leq \langle DA(u_N(\mu); \mu)(H(z^2; \mu) - H(z^1; \mu)), T_N^\mu (H(z^2; \mu) - H(z^1; \mu)) \rangle.
\end{equation}

Let $\alpha > 0$; if $z_1$ and $z_2$ are in $B_{\overline{\Pi}}(u_N(\mu), \alpha)$, then $\| u_N(\mu) - \frac{1}{2}(z^2 + z^1) \|_{\overline{\Pi}} < \alpha$, and the last two inequalities give

\[
\| H(z^2; \mu) - H(z^1; \mu) \|_{\overline{\Pi}} < \frac{\beta_{\overline{\Pi}}(\mu) \alpha}{\beta_{\overline{\Pi}}(\mu)} \| z^2 - z^1 \|_{\overline{\Pi}},
\]

which implies that $H(\cdot; \mu)$ is a contraction mapping for

\[
\alpha < \frac{\beta_{\overline{\Pi}}(\mu)}{\rho_{\overline{\Pi}}(\mu)}.
\]

It follows that there can be only (at most) one fixed point of $H(\cdot; \mu)$ inside $B_{\overline{\Pi}}(u_N(\mu), \beta_{\overline{\Pi}}(\mu))$ and hence at most one solution $u_N(\mu)$. \hfill \square

3.2. Existence. We define the following quantities:

\begin{align}
\epsilon_N(\mu) &\equiv \| A(u_N(\mu); \mu) \|_X, \\
\tau_{N, \overline{\Pi}}(\mu) &\equiv \frac{2\epsilon_N(\mu)\epsilon_N(\mu)}{\beta_{\overline{\Pi}}(\mu)^2}, \\
\Delta_{N, \overline{\Pi}}(\mu) &\equiv \frac{\beta_{\overline{\Pi}}(\mu)}{\rho_{\overline{\Pi}}(\mu)} \left[ 1 - \sqrt{1 - \tau_{N, \overline{\Pi}}(\mu)} \right].
\end{align}
Lemma 3.2. If $\beta_{\overline{\mu}}(\mu) > 0$ and $\tau_{N,\overline{\mu}}(\mu) \leq 1$, there exists a unique solution $u(\mu)$ to (2.1) such that

$$u(\mu) \in B_{\overline{\mu}}(u_N(\mu), \Delta_{N,\overline{\mu}}(\mu)).$$

Proof. Let $\alpha > 0$ and $z$ in $B_{\overline{\mu}}(u_N(\mu), \alpha)$; consider

$$H(z; \mu) - u_N(\mu) = z - u_N(\mu) - DA(u_N(\mu); \mu)^{-1}(A(z; \mu) - A(u_N(\mu); \mu)) - DA(u_N(\mu); \mu)^{-1}A(u_N(\mu); \mu).$$

By using arguments similar to those invoked in the previous lemma (among others, that $u(\mu)$ is a fixed point of $H(\cdot; \mu)$), we have

$$\beta_{\overline{\mu}}(\mu)\|H(z; \mu) - u_N(\mu)\|_\overline{\mu}\|T^N_{\mu}(H(z; \mu) - u_N(\mu))\|_X$$

for all $v$ in $X$. Similarly to (3.7),

$$\beta_{\overline{\mu}}(\mu)\|H(z; \mu) - u_N(\mu)\|_\overline{\mu}\|T^N_{\mu}(H(z; \mu) - u_N(\mu))\|_X$$

Since $z$ is in the open ball $B_{\overline{\mu}}(u_N(\mu), \alpha)$, choosing $v = T^N_{\mu}(H(z; \mu) - u_N(\mu))$, (3.11), and (3.12) gives

$$\|H(z; \mu) - u_N(\mu)\|_\overline{\mu} \leq \frac{\rho_{\overline{\mu}}(\mu)}{2\beta_{\overline{\mu}}(\mu)}\|z - u_N(\mu)\|_\overline{\mu}^2 + \frac{\epsilon_{N,\overline{\mu}}(\mu)}{\beta_{\overline{\mu}}(\mu)} < \frac{\rho_{\overline{\mu}}(\mu)\alpha^2}{2\beta_{\overline{\mu}}(\mu)} + \frac{\epsilon_{N,\overline{\mu}}(\mu)}{\beta_{\overline{\mu}}(\mu)}.$$

We now look for values of $\alpha$ (in particular, the smallest will give us a sharper error bound) such that

$$\frac{\rho_{\overline{\mu}}(\mu)\alpha^2}{2\beta_{\overline{\mu}}(\mu)} + \frac{\epsilon_{N,\overline{\mu}}(\mu)}{\beta_{\overline{\mu}}(\mu)} \leq \alpha.$$

Provided that $\tau_{N,\overline{\mu}}(\mu) \leq 1$, we have that the smallest value is given by $\alpha = \Delta_{N,\overline{\mu}}(\mu)$. Then for this choice and from (3.13) we have

$$\|H(z; \mu) - u_N(\mu)\|_\overline{\mu} < \alpha;$$

i.e., $H$ maps $B_{\overline{\mu}}(u_N(\mu), \Delta_{N,\overline{\mu}}(\mu))$ into itself. The proof follows from $\Delta_{N,\overline{\mu}}(\mu) \leq \frac{\beta_{\overline{\mu}}(\mu)}{\rho_{\overline{\mu}}(\mu)}$, the fact that $H$ is a contraction (cf. the proof of the previous lemma), and the contraction mapping theorem.

3.3. A posteriori error estimation.

Theorem 3.3. Let $\mu \in D$, and assume that, for a $\overline{\mu} \in V^K$, $\beta_{\overline{\mu}}(\mu) > 0$ and $\tau_{\overline{\mu}}(\mu) \leq 1$. Then there exists a unique solution $u(\mu)$ such that

$$\|u(\mu) - u_N(\mu)\|_\overline{\mu} < \frac{\beta_{\overline{\mu}}(\mu)}{\rho_{\overline{\mu}}(\mu)};$$
the error with respect to the natural norm is

\begin{equation}
\|u(\mu) - u_N(\mu)\|_{\mathcal{P}} \leq \Delta_{N,\mathcal{P}}(\mu)
\end{equation}

with effectivity

\begin{equation}
\Delta_{N,\mathcal{P}}(\mu) \leq \left[ \frac{2\gamma_{\mathcal{P}}(\mu)}{\beta_{\mathcal{P}}(\mu)} + \tau_{\mathcal{P}}(\mu) \right] \|u(\mu) - u_N(\mu)\|_{\mathcal{P}}.
\end{equation}

This theorem provides an error bound with respect to the natural norms that, compared to previous results (cf. [17, 33]), first provides better effectivity in the field variable, since we expect \(\gamma_{\mathcal{P}}(\mu)\) and \(\beta_{\mathcal{P}}(\mu)\) to be of order one and second indirectly sharpens the existence’s condition since in (3.9) again \(\beta_{\mathcal{P}}(\mu)\) is of order one.

Note that we expect that \(\gamma_{\mathcal{P}}(\mu)\) and \(\beta_{\mathcal{P}}(\mu)\) are of order one, and hence \(\mathcal{T}_N^\infty\) and \(\mathcal{T}^\infty\) have similar spectra. In (3.9) the term that is the worse controlled is \(\rho_{\mathcal{P}}(\mu)\); this is related to a Sobolev embedding constant (cf. [33] and section 6). However, the equivalent condition in [33] reads

\[ \tau_N(\mu) = \frac{2\rho \epsilon_N(\mu)}{\beta_N(\mu)^2}, \]

where \(\rho\) and \(\beta_N(\mu)\) are the Sobolev embedding constants and equivalent to (2.8) and (3.1), respectively, with respect to the \(X\)-norm only. Since \(\beta_{\mathcal{P}}(\mu) \geq \frac{\beta_N(\mu)}{\gamma_{\mathcal{P}}(\mu)}\) and \(\rho_{\mathcal{P}}(\mu) \leq \frac{\rho}{\beta_N(\mu)}\), we have

\[ \tau_{N,\mathcal{P}}(\mu) \leq \tau_N(\mu) \frac{\gamma(\mu)^2}{\beta(\mu)^2}; \]

i.e., if \(\gamma_{\mathcal{P}}(\mu)\) and \(\beta_{\mathcal{P}}(\mu)\) are of order one, the condition \(\tau_{N,\mathcal{P}}(\mu) \leq 1\) is less stringent than \(\tau_N(\mu) < 1\), which is required in [33].

In our simulations, for a large Grashof number, we actually get that \(\rho_{\mathcal{P}}(\mu)\) is one order smaller than \(\frac{\rho}{\beta_N(\mu)}\), and hence \(\tau_{N,\mathcal{P}}(\mu)\) is also one order smaller than \(\tau_N(\mu)\).

Proof. We just need to prove (3.15). Let \(e(\mu) = u(\mu) - u_N(\mu)\) and \(\hat{e}(\mu)\) such that

\begin{equation}
(\hat{e}(\mu), v)_X = -\langle A(u_N(\mu); \mu), v \rangle
= f(v) - a(u_N(\mu), v; \mu) = a(u(\mu), v; \mu) - a(u_N(\mu), v; \mu).
\end{equation}

Note that, from (3.8), \(\|\hat{e}(\mu)\|_X = \epsilon_{N,\mathcal{P}}(\mu)\) and that, since our nonlinearity is quadratic,

\[ a(u(\mu), v; \mu) - a(u_N(\mu), v; \mu) = da(u_N(\mu); \mu)(e(\mu), v) + \frac{1}{2} \left( da(u(\mu); \mu) - da(u_N(\mu); \mu) \right)(e(\mu), v) \]

for all \(v\) in \(X\). This, (3.16), (3.2), and (2.8) yield to

\[ \|\hat{e}(\mu)\|_X^2 = da(u_N(\mu); \mu)(e(\mu), \hat{e}(\mu)) + \frac{1}{2} \left( da(u(\mu); \mu) - da(u_N(\mu); \mu) \right)(e(\mu), \hat{e}(\mu)) \]

\[ \leq \gamma_{\mathcal{P}}(\mu)\|e(\mu)\|_{\mathcal{P}}\|\hat{e}(\mu)\|_X + \frac{1}{2} \rho_{\mathcal{P}}(\mu)\|e(\mu)\|_{\mathcal{P}}^2 \|\hat{e}(\mu)\|_X \]
and therefore

\[ (3.17) \quad \epsilon_{N,\pi}(\mu) \leq \gamma_{\pi}(\mu) \| e(\mu) \|_{\pi} + \frac{1}{2} \mu_{\pi}(\mu) \| e(\mu) \|_{\pi}^2. \]

Since \( 0 \leq \tau_{N,\pi}(\mu) \leq 1 \) and \( 1 - \sqrt{1 - \tau_{N,\pi}(\mu)} \leq \tau_{N,\pi}(\mu) \), from (3.9)–(3.10) we get

\[ (3.18) \quad \Delta_{N,\pi}(\mu) \leq 2 \epsilon_{N,\pi}(\mu) \frac{\mu_{\pi}(\mu)}{\beta_{\pi}(\mu)} \quad \text{and} \]

\[ (3.19) \quad \frac{\mu_{\pi}(\mu)}{\beta_{\pi}(\mu)} \Delta_{N,\pi}(\mu) \leq \tau_{N,\pi}(\mu). \]

Then, from (3.17) and (3.18)

\[ \Delta_{N,\pi}(\mu) \leq 2 \epsilon_{N,\pi}(\mu) \frac{\mu_{\pi}(\mu)}{\beta_{\pi}(\mu)} \| e(\mu) \|_{\pi} + \frac{\mu_{\pi}(\mu)}{\beta_{\pi}(\mu)} \| e(\mu) \|_{\pi}^2; \]

we split the square of \( \| e(\mu) \|_{\pi} \) and get

\[ \Delta_{N,\pi}(\mu) \leq 2 \epsilon_{N,\pi}(\mu) \frac{\mu_{\pi}(\mu)}{\beta_{\pi}(\mu)} \| e(\mu) \|_{\pi} + \frac{\mu_{\pi}(\mu)}{\beta_{\pi}(\mu)} \| e(\mu) \|_{\pi} \Delta_{N,\pi}(\mu) \]

\[ \leq 2 \epsilon_{N,\pi}(\mu) \frac{\mu_{\pi}(\mu)}{\beta_{\pi}(\mu)} \| e(\mu) \|_{\pi} + \tau_{N,\pi}(\mu) \| e(\mu) \|_{\pi}, \]

thanks to (3.14) in the first inequality and (3.19) in the second one. \( \Box \)

3.4. Output error bound. Often the precision required in the prediction of the output is obtained by the primal approximation alone. This is the case in the model problems that we propose; in fact, the requirements to ensure existence and unicity already imply a very small error in the output.

For a given \( \mu \in D \), we are interested in bounding the error on the output \( s(\mu) = \ell(u(\mu)) \). For \( \pi \in V^K \), the natural dual norm of the output linear functional is defined as

\[ \| \ell \|_{\pi'} = \sup_{v \in Y} \left| \frac{\ell(v)}{\| v \|_{\pi}} \right|. \]

(Note that, in the case of an affine output, \( \| \ell \|_{\pi'} \leq \sum_{q=1}^{Q_i} |\theta_q(\mu)| \| e_q(\mu) \|_{\pi'} \).) If \( \beta_{\pi}(\mu) > 0 \) and \( \tau_{N,\pi}(\mu) \leq 1 \), we can then state the output error bound as

\[ |s(\mu) - s_N(\mu)| \leq \| \ell \|_{\pi} \Delta_{N,\pi}(\mu) \equiv \Delta^*_N(\mu). \]

However, we do not have any effectivity result about these bounds. In our example, the effectivities are at high values of the Grashof number reasonable but not as accurate at low values. In the appendix we propose a dual approach that partially addresses this issue; further improvements shall require deflation (cf. [32]).

4. Lower bound of the inf-sup constant. We are going to build a lower bound for the inf-sup parameter \( \beta_{\pi}(\mu) \). The ingredients, similarly to [32], are the selection of a candidate supremizer, a Taylor expansion with the remainder, and a lower bound for the terms in the Taylor expansion.
It is clear that, for $\mu = \overline{\mu}$ in (3.1) and (3.2), $\beta_{\overline{\mu}}(\mu) = \gamma_{\overline{\mu}}(\mu) = 1$; the natural norm can thus be viewed as a generalization of the usual energy norm (for symmetric, coercive operators) to the nonlinear case. It can be demonstrated that

$$\beta(\mu)/\gamma_{\overline{\mu}}(\mu) \leq \beta(\mu) \leq \gamma_{\overline{\mu}}(\mu)\gamma(\mu).$$

In what follows we shall require an “intermediate” inf-sup parameter—an approximation to $\beta_{\overline{\mu}}(\mu)$—which we denote as $\beta_{\overline{\mu}}(\mu)$: For given $\mu \in \mathcal{D}$ and $\overline{\mu} \in V^R$,

$$\beta_{\overline{\mu}}(\mu) \equiv \inf_{\nu \in \mathcal{X}} \frac{(T^\mu_N \nu - T^\mu_N \overline{\mu}, T^\mu_N \nu + T^\mu_N \overline{\mu})_X}{\|T^\mu_N \nu\|_X^2}.$$

It follows directly from the Cauchy–Schwarz inequality—or, equivalently, we may observe that $T^\mu_N \nu$ is a candidate supremizer $\nu$ in (3.1)—that

$$\beta_{\overline{\mu}}(\mu) \leq \beta(\mu) \quad \forall \mu \in \mathcal{D}.$$  

(Note that $\beta_{\overline{\mu}}(\mu)$ is not necessarily positive.)

We can also show that $\beta_{\overline{\mu}}(\mu)$ is a “good” lower bound for $\beta_{\overline{\mu}}(\mu)$.

**Lemma 4.1.** Let

$$\delta^\mu_{\overline{\mu}} = \sup_{\nu \in \mathcal{X}} \frac{(T^\mu_N \nu - T^\mu_N \overline{\mu}, T^\mu_N \nu + T^\mu_N \overline{\mu})_X}{\|T^\mu_N \nu\|_X^2},$$

and assume that

$$\sup_{\nu \in \mathcal{X}} \frac{\|T^\mu_N \nu - T^\mu_N \overline{\mu}\|_X}{\|T^\mu_N \overline{\mu}\|_X} = \delta(\delta^\mu_{\overline{\mu}})$$

as $\mu \to \overline{\mu}$. Then

$$1 - \beta_{\overline{\mu}}(\mu) = -\frac{1}{2} \delta^\mu_{\overline{\mu}} + o(\|\delta^\mu_{\overline{\mu}}\|^2),$$

and there exists $c(\overline{\mu}, \mu)$ in [0, 1] such that

$$c(\overline{\mu}, \mu) [1 - \beta_{\overline{\mu}}(\mu)] + 2 [\beta_{\overline{\mu}}(\mu) - \beta_{\overline{\mu}}(\mu)] = \delta(\delta^\mu_{\overline{\mu}}^2).$$

In general, we can assume that $\delta^\mu_{\overline{\mu}}$ is of order $o(\|\mu - \overline{\mu}\|)$; hence, as for the linear case [32], $1 - \beta_{\overline{\mu}}(\mu) = o(\|\mu - \overline{\mu}\|)$ as $\mu \to \overline{\mu}$ and

$$0 \leq \beta_{\overline{\mu}}(\mu) - \beta_{\overline{\mu}}(\mu) \leq \frac{1}{2} [1 - \beta_{\overline{\mu}}(\mu)] + o(\|\overline{\mu} - \mu\|^2).$$

**Proof.** From the definition of $\beta_{\overline{\mu}}(\mu)$ and $\beta_{\overline{\mu}}(\mu)$, $1 - \beta_{\overline{\mu}}(\mu)^2 = -\delta^\mu_{\overline{\mu}}$ and $\beta_{\overline{\mu}}(\mu)^2 \leq \beta_{\overline{\mu}}(\mu)^2 \leq \beta_{\overline{\mu}}(\mu)^2 + \frac{1}{2} \delta^\mu_{\overline{\mu}} + o(\|\delta^\mu_{\overline{\mu}}\|^2)$; hence

$$\beta_{\overline{\mu}}(\mu)^2 - \beta_{\overline{\mu}}(\mu)^2 = -\frac{c(\overline{\mu}, \mu)}{2} \delta^\mu_{\overline{\mu}} + o(\|\delta^\mu_{\overline{\mu}}\|^2),$$

where $c(\overline{\mu}, \mu) \in [0, 1]$. From the Taylor series of the square root follows (4.1) and

$$\beta_{\overline{\mu}}(\mu) - \beta_{\overline{\mu}}(\mu) = \frac{c(\overline{\mu}, \mu)}{4} \delta^\mu_{\overline{\mu}} + o(\|\delta^\mu_{\overline{\mu}}\|^2).$$

Since $\frac{1}{1-x} = 1 + x + o(x)$, and from (4.1), $\beta_{\overline{\mu}}(\mu) - \beta_{\overline{\mu}}(\mu) = \frac{c(\overline{\mu}, \mu)}{4} \delta^\mu_{\overline{\mu}} + o(\|\delta^\mu_{\overline{\mu}}\|^2)$, which proves (4.2). \qed
4.1. Local lower bound. For a fixed \( \overline{\mu} \) in \( V^K \), we expand \( \overline{\beta}(\mu) \) as follows:

\begin{align}
\overline{\beta}(\mu) &= \inf_{w \in X} \left( \frac{\sum_{q=1}^{Q_0} a_0^q(w, T^f_N w)}{\|w\|^2} \right) \\
&= \inf_{w \in X} \left\{ 1 + \frac{1}{\|w\|^2} \sum_{q=1}^{Q_1} (\Theta_0^q(\mu) - \Theta_0^q(\overline{\mu})) a_0^q(w, T^f_N w) \\
&\quad + \frac{1}{\|w\|^2} \sum_{q=1}^{Q_1} (\Theta_1^q(\mu) - \Theta_1^q(\overline{\mu})) a_1^q(w, u_N(\mu), T^f_N w) \\
&\quad + \frac{1}{\|w\|^2} \sum_{q=1}^{Q_1} (\Theta_1^q(\mu) - \Theta_1^q(\overline{\mu})) a_1^q(w, u_N(\mu) - u_N(\overline{\mu}), T^f_N w) \right\}.
\end{align}

We approximate the partial derivative of \( \frac{\partial u_N}{\partial \mu_p} \) at \( \overline{\mu} \) with respect to the \( p \)th parameter in our reduced space \( W_N \):

\begin{align}
\frac{\partial u_N}{\partial \mu_p}(\overline{\mu}) &= \left( \frac{\partial u_N}{\partial \mu_p}(\overline{\mu}), v \right) \\
&= - \left[ \sum_{q=1}^{Q_1} \frac{\partial \Theta_0^q(\overline{\mu})}{\partial \mu_p} a_0^q(u_N(\overline{\mu}), v) + \frac{1}{2} \sum_{q=1}^{Q_1} \frac{\partial \Theta_1^q(\overline{\mu})}{\partial \mu_p} a_1^q(u_N(\overline{\mu}), u_N(\overline{\mu}), v) \right]
\end{align}

for all \( v \) in \( W_N \). (If \( f \) affinely depends on the parameter, we have to add \( \sum_{q=1}^{Q_1} \frac{\partial \Theta_1^q(\overline{\mu})}{\partial \mu_p} a_1^q(u_N(\overline{\mu}), v) \) to the right-hand side of (4.4).) We then rewrite

\begin{align}
a_1^q(w, u_N(\mu) - u_N(\overline{\mu}), T^f_N w) \\
&= \sum_{p=1}^{P} \kappa_p (\mu_p - \overline{\mu}_p) a_1^q(w, \frac{\partial u_N}{\partial \mu_p}(\overline{\mu}), T^f_N w) + a_1^q(w, \alpha(\mu, \overline{\mu}, \kappa), T^f_N w),
\end{align}

where \( \kappa \in \mathbb{R}^P \) and \( \alpha(\mu, \overline{\mu}, \kappa) \in W_{N_{\max}} \) is defined as

\begin{align}
\alpha(\mu, \overline{\mu}, \kappa) = u_N(\mu) - u_N(\overline{\mu}) - \sum_{p=1}^{P} \kappa_p (\mu_p - \overline{\mu}_p) \frac{\partial u_N}{\partial \mu_p}(\overline{\mu}).
\end{align}

As for the linear case [32], we perform a Taylor expansion with respect to \( \mu \) at \( \overline{\mu} \). The first-order terms in (4.3a), (4.3b), and (4.3c) are collected: For \( p = 1, \ldots, P \),

\begin{align}
a^p(\cdot, \cdot) &= \sum_{q=1}^{Q_0} \frac{\partial \Theta_0^q(\overline{\mu})}{\partial \mu_p} a_0^q(\cdot, \cdot) + \sum_{q=1}^{Q_1} \frac{\partial \Theta_1^q(\overline{\mu})}{\partial \mu_p} a_1^q(\cdot, u_N(\overline{\mu}), \cdot) + \sum_{q=1}^{Q_1} \Theta_1^q(\overline{\mu}) a_1^q(\cdot, \frac{\partial u_N}{\partial \mu_p}(\overline{\mu}), \cdot).
\end{align}
We denote the extreme eigenvalues with respect to $\|\cdot\|_\mathcal{P}$ of $\overline{a}_p^\mathcal{P}, a_0^\mathcal{P}$, and $a_1^\mathcal{P}(\cdot, u_N(\overline{\mu}), \cdot)$ as
\[
\lambda_{p, \inf}^\mathcal{P} = \inf_{w \in \mathcal{X}} \frac{a_p^\mathcal{P}(w, T_N^\mathcal{P} w)}{\|w\|^2_{\mathcal{P}}}, \quad \lambda_{p, \sup}^\mathcal{P} = \sup_{w \in \mathcal{X}} \frac{a_p^\mathcal{P}(w, T_N^\mathcal{P} w)}{\|w\|^2_{\mathcal{P}}}, \quad p = 1, \ldots, P,
\]
\[
\gamma_{0,q, \inf}^\mathcal{P} = \inf_{w \in \mathcal{X}} \frac{a_0^\mathcal{P}(w, T_N^\mathcal{P} w)}{\|w\|^2_{\mathcal{P}}}, \quad \gamma_{0,q, \sup}^\mathcal{P} = \sup_{w \in \mathcal{X}} \frac{a_0^\mathcal{P}(w, T_N^\mathcal{P} w)}{\|w\|^2_{\mathcal{P}}}, \quad q = 1, \ldots, Q_0,
\]
\[
\gamma_{1,q, \inf}^\mathcal{P} = \inf_{w \in \mathcal{X}} \frac{a_1^\mathcal{P}(w, u_N(\overline{\mu}), T_N^\mathcal{P} w)}{\|w\|^2_{\mathcal{P}}}, \quad q = 1, \ldots, Q_1.
\]

From our assumption (2.9) we have
\[
(4.6) \quad \frac{1}{\|w\|_{\mathcal{P}}} \left| \sum_{q=1}^{Q_1} \Theta^q_1(\mu) a_1^\mathcal{P}(w, \alpha(\mu, \overline{\mu}, \kappa), T_N^\mathcal{P} w) \right| \leq \rho_{\mathcal{X}, \mathcal{P}} \|\alpha(\mu, \overline{\mu}, \kappa)\|_{\mathcal{X}}.
\]

Similarly, from (2.10),
\[
\frac{1}{\|w\|_{\mathcal{P}}} \left| \sum_{q=1}^{Q_1} (\Theta^q_1(\mu) - \Theta^q_1(\overline{\mu})) a_1^\mathcal{P}(w, u_N(\mu) - u_N(\overline{\mu}), T_N^\mathcal{P} w) \right| \leq \max_{q=1, \ldots, Q_1} \frac{|\Theta^q_1(\mu) - \Theta^q_1(\overline{\mu})|}{|\Theta^q_1(\overline{\mu})|} \rho_{\mathcal{X}, \mathcal{P}} \|u_N(\mu) - u_N(\overline{\mu})\|_{\mathcal{X}}.
\]

As a result, our bound to (4.3) is defined, for any $\kappa$ in $\mathbb{R}^P$, as
\[
(4.7) \quad \beta_{\mathcal{P}}^L(\mu, \kappa) \equiv 1 + \sum_{p=1}^{P} \min_{\lambda_p = \lambda_{p, \inf}^\mathcal{P}, \sup} \kappa_p(\mu_p - \overline{\mu}_p) \lambda_p
\]
\[
+ \sum_{q=1}^{Q_0} \min_{\gamma_q = \gamma_{0,q, \inf}, \sup} \left( \Theta^q_0(\mu) - \Theta^q_0(\overline{\mu}) - \sum_{p=1}^{P} \kappa_p(\mu_p - \overline{\mu}_p) \frac{\partial \Theta^q_0}{\partial \mu_p}(\overline{\mu}) \right) \gamma_q
\]
\[
+ \sum_{q=1}^{Q_1} \min_{\gamma_q = \gamma_{1,q, \inf}, \sup} \left( \Theta^q_1(\mu) - \Theta^q_1(\overline{\mu}) - \sum_{p=1}^{P} \kappa_p(\mu_p - \overline{\mu}_p) \frac{\partial \Theta^q_1}{\partial \mu_p}(\overline{\mu}) \right) \gamma_q
\]
\[
- \rho_{\mathcal{X}, \mathcal{P}} \left( \|\alpha(\mu, \overline{\mu}, \kappa)\|_{\mathcal{X}} + \max_{q=1, \ldots, Q_1} \frac{|\Theta^q_1(\mu) - \Theta^q_1(\overline{\mu})|}{|\Theta^q_1(\overline{\mu})|} \|u_N(\mu) - u_N(\overline{\mu})\|_{\mathcal{X}} \right),
\]

where we denote the minimum of a function $g$ over two values as
\[
\min_{\lambda = \lambda_{\inf}, \sup} g(\lambda) \equiv \min \{g(\lambda_{\inf}), g(\lambda_{\sup})\}.
\]

Since (4.7) is valid for any $\kappa \in \mathbb{R}^P$, we define our inf-sup lower bound as
\[
(4.8) \quad \beta_{\mathcal{P}}^{LB}(\mu, \kappa) \equiv \max_{\kappa \in \mathbb{R}^P} \beta_{\mathcal{P}}^{LB}(\mu, \kappa).
\]

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.
4.2. \( \kappa \) optimization. Ideally, for a given \( \mu \), we have to find the value of \( \kappa \) that maximizes \( B_{\mathbf{p}}^{LB}(\mu, \kappa) \). The optimization of \( \kappa \) relies on quadratic programming. We propose a simpler heuristic method: The bound involving the eigenvalues \( \lambda_{p, \inf|sup} \) is sharper than (4.6), and therefore we compute \( \kappa \) such that \( \| \alpha(\mu, \mathbf{\bar{p}}, \kappa) \|_X \) is the smallest. Moreover, in our algorithms we will require an \( X \)-orthonormal basis for \( W_{N_{\max}} \); hence this minimization is achieved by the least-square solution \( \kappa^* \) that satisfies

\[
\begin{bmatrix}
  \vdots \\
  (\mu_p - \mathbf{\bar{p}}_p) \partial_p U_{N,j}(\mathbf{\bar{p}}) \\
  \vdots \\
  \vdots \\
  j,p \\
\end{bmatrix}^T
\begin{bmatrix}
  \vdots \\
  (\mu_p - \mathbf{\bar{p}}_p) \partial_p U_{N,j}(\mathbf{\bar{p}}) \\
  \vdots \\
  \vdots \\
  j,p \\
\end{bmatrix}
\kappa^* =
\begin{bmatrix}
  \vdots \\
  (\mu_p - \mathbf{\bar{p}}_p) \partial_p U_{N,j}(\mathbf{\bar{p}}) \\
  \vdots \\
  \vdots \\
  j,p \\
\end{bmatrix}^T
\begin{bmatrix}
  \vdots \\
  U_{N,j}(\mu) - U_{N,j}(\mathbf{\bar{p}}) \\
  \vdots \\
  \vdots \\
  j \\
\end{bmatrix},
\]

where \( U_{N,j}(\mu) \), \( U_{N,j}(\mathbf{\bar{p}}) \), and \( \partial_p U_{N,j}(\mathbf{\bar{p}}) \) are the \( j \)th components of \( u_N(\mu) \), \( u_N(\mathbf{\bar{p}}) \), and \( \partial_p u_N(\mu) \) with respect to the basis of \( W_{N_{\max}} \).

5. Computational strategy.

5.1. Offline algorithm. Our offline algorithm includes the computation of the reduced basis ingredients—optimal selection of the basis and computation of matrices and vector in the reduced space (cf., e.g., [32])—and of the a posteriori error estimation ingredients—selection of \( Y^K = \{ \mathbf{\bar{p}}_1, \ldots, \mathbf{\bar{p}}_K \} \), solution of the eigenvalue problems, and computation of the Sobolev embedding constants. Because of the presence of the reduced basis approximations \( u_N(\mathbf{\bar{p}}) \) in the definition of the natural norms, these two stages are dependent on each other. (In the linear case (cf. [32]) these steps are ordered: first to compute the a posteriori error estimation ingredients, then to select the reduced basis.)

In the first step, we provide surrogates \( B_{\mathbf{p}}^{LB}(\mu) \) and \( \rho(\mathbf{\bar{p}}) \) in order to compute an error bound extrapolation. We expect that \( B_{\mathbf{p}}^{LB}(\mu) \) is of order one, so we set its surrogate to, say, 0.2; to compute a surrogate \( \rho(\mathbf{\bar{p}}) \), we replace \( u_N(\mathbf{\bar{p}}) \) by \( u(\mathbf{\bar{p}}) \) in (2.4) (hence in (2.7))—in other words we do not need a reduced basis. As a result we have an efficient—online fast—tool to approximately compute the error for a given \( \mu \) in \( D \).

In the second step, we compute the eigenvalues and the Sobolev embedding constants that we need in our inf-sup lower bound construction.

We hence propose the following algorithm:

1. Manually set \( K \), select a representative set \( Y^K = \{ \mathbf{\bar{p}}_1, \ldots, \mathbf{\bar{p}}_K \} \), and compute surrogates for each \( \rho(\mathbf{\bar{p}}) \);
2. start an optimal search algorithm: For a large random\(^{5}\) set of \( \mu \)'s
   (a) solve the reduced basis problem with the existing basis;
   (b) select the “nearest” \( \mathbf{\bar{p}} \) in \( Y^K \), and compute \( \epsilon_N(\mu), \tau_N(\mathbf{\bar{p}}, \mu), \) and \( \Delta_{N,\mathbf{\bar{p}}}(\mu) \) by replacing \( \rho(\mathbf{\bar{p}}) \) and \( \beta(\mu) \) by their surrogates;
   (c) select the optimal \( \mu \) as follows:

\(^{5}\)This can be replaced by a deterministic search over a large set of sample points.
i. If for some \( \mu \)'s, (2a) did not converge, select the “furthest” from the previously selected ones;

ii. if for some \( \mu \)'s, \( \tau_{N,\bar{\pi}}(\mu) \) is bigger than 1, select the one with the largest \( \tau_{N,\bar{\pi}}(\mu) \);

iii. otherwise, select the one with the largest \( \Delta_{N,\bar{\pi}}(\mu) \).

3. Solve the finite element problem (2.1) for the selected \( \mu \);

4. increase \( N \) and enrich the reduced basis space \( W_N \) with \( u(\mu) \);

5. if the maximum number of bases wished or the tolerance requested for \( \Delta_{N,\bar{\pi}}(\mu) \) are achieved, go to 6; otherwise, go to 2;

6. set \( N \) to the reached number of basis functions, and compute all of the ingredients for the computation of \( \mathbf{B}_{LB}(\mu) \);

7. start a random process: For a large set of \( \mu \)'s check that the reduced basis problem can be solved, that \( \mathbf{B}_{LB}(\mu) \) is positive for at least one \( \mu \), and that \( \tau_{N,\bar{\pi}}(\mu) \leq 1 \) for this \( \mu \).

   If this is ok, then all of the online components are ready, and we set \( N_{\text{max}} = N \); otherwise, either add some \( \mu \)'s to \( V_K \) and go to 6 or go to 2 and replace \( \beta_{\bar{\pi}}(\mu) \) by \( \mathbf{B}_{LB}(\mu) \) instead of by 0.2.

   Steps 2(a) and 2(b) are online fast. Moreover, in our parallel offline code, each processor performs the optimal search on a different random set. As a result, we can explore our parameter space very rapidly.

To ensure that we are on a selected solution branch, in the finite element resolution we perform a homotopy procedure in \( \mu \) starting from an already computed solution at a nearby parameter; in the reduced basis we simply start from a solution at a nearby parameter used to generate the \( W_N \). (In some applications, it may be necessary to perform homotopy also in the reduced basis context.)

Note that we can admit small errors in the resolution of (2.1); indeed, we need approximations to \( u(\mu) \) at selected parameter points to build our reduced basis. However, the error bound is given with respect to the exact “truth” solution \( u(\mu) \).

The reader is referred to [33] and [32] for the exact offline procedure to compute the reduced basis and a posteriori error estimation components.

### 5.2. Online algorithm

The online procedure, for given \( \mu \in D \) and \( N \leq N_{\text{max}} \), reads:

I. Solve the reduced basis problem (1.3)\(^8\) and compute \( s_N(\mu) = \ell(u(\mu)) \);

II. compute \( \epsilon_N(\mu) \);

III. order \( V_K \) in increasing distance\(^9\) from \( \mu \). Compute \( \mathbf{B}_{LB}(\mu) \) and \( \tau_{N,\bar{\pi}}(\mu) \) until we find \( \bar{\mu} \) in \( V \) such that \( \mathbf{B}_{LB}(\mu) > 0 \) and \( \tau_{N,\bar{\pi}}(\mu) \leq 1 \). We set \( I_K(\mu) = \bar{\mu} \);

IV. compute \( \Delta_{N,\bar{\pi}}(\mu) \) and \( \Delta_{N,\bar{\pi}}(\mu) \).

If in step III we do not find any \( \bar{\mu} \), our method fails: We cannot provide an error bound nor even existence; if possible, \( N \) must be increased; otherwise, our basis is not rich enough.

---

\(^6\)This step involves Gramm–Schmidt orthogonalization and the computation of the matrix- and scalar-products necessary for the reduced basis matrices and the online computation of the Y-dual norm of the residual (cf. [23, 32]).

\(^7\)In our model problems, \( u(\mu) \) is split into its physical components and a stabilizing supremizer is computed; the reduced space is hence enriched by four functions (cf. section 6.2).

\(^8\)We solve this by a Newton algorithm with the initial guess a solution at a nearby parameter used to generate the reduced basis.

\(^9\)The distance function in \( D \) is arbitrary.
The computational effort is independent of $N$: Steps I and II have dominant complexities $N^3 + Q_1 N^2$ times the number of Newton iterations required and $Q_1^2 N^4 + Q_0 Q_1 N^3$, respectively. The complexity of steps III and IV is dominated by that of step II.

6. Application: Steady natural convection. We are interested in solving a natural convection problem in a two-dimensional square cavity under vertical gravity (Figure 1), with constant temperature on one side, heated on the opposite side, insulated at the top and bottom, and Prandtl number $\Pr = 1$ (dissipation to conduction ratio). We consider one parameter $\mu = (1/\sqrt{\text{Gr}})$, where Gr is the Grashof number, which is defined as $(\beta \Delta T g L^3)/\nu^2$ ($g$ is the gravity constant, $\beta$ is the thermal expansion coefficient, $\Delta T$ is the temperature difference, $L$ is the length scale, and $\nu$ is the kinematic viscosity).

We impose homogeneous Dirichlet boundary conditions for the velocity $(u_1, u_2)$, while for the temperature $\theta$ we impose homogeneous Dirichlet boundary conditions on the left side, homogeneous Neumann boundary conditions on the top and the bottom, and constant unitary Neumann boundary conditions on the heated side $\Gamma_h$ ($\partial_n \theta = 1$ on $\Gamma_h$). We scale the pressure such that $\int_{\Omega} p = 0$.

Our domain $\Omega \subset \mathbb{R}^2$ is a closed square cavity with unitary sides. We want to solve (with summation over repeated indices and partial derivatives with respect to $x_1$ and $x_2$)

\[
\begin{align*}
    u_j \partial_j u_1 &= -\partial_1 p + \frac{1}{\sqrt{\text{Gr}}} \partial_{jj}^2 u_1, \\
    u_j \partial_j u_2 &= -\partial_2 p + \theta + \frac{1}{\sqrt{\text{Gr}}} \partial_{jj}^2 u_2, \\
    \partial_i u_i &= 0, \\
    u_j \partial_j \theta &= \frac{1}{\sqrt{\text{Gr} \Pr}} \partial_{jj}^2 \theta
\end{align*}
\]

in $\Omega$.

Our outputs of interest are the inverse of the Nusselt number and the flux through the middle half section $\Gamma_s$ parallel to $\Gamma_h$:

\[
\begin{align*}
    s_{(1)}(\mu) &= \frac{1}{|\Gamma_h|} \int_{\Gamma_h} \theta(\mu), \\
    s_{(2)}(\mu) &= \frac{1}{|\Gamma_s|} \int_{\Gamma_s} u_1(\mu).
\end{align*}
\]
6.1. Affine decomposition. Our reference finite element space \( X \) is a subset of \( H^1_0(\Omega) \times H^1_0(\Omega) \times H^1(\Omega) \times L^2(\Omega) \times \mathbb{R} \). We denote elements in \( X \) as \( \mathbf{u} = (u_1, u_2, \theta, p, \lambda) \), \( \mathbf{v} = (v_1, v_2, \chi, q, \gamma) \), and \( \mathbf{z} = (z_1, z_2, \zeta, \ldots) \).

Recalling (2.2), \( \sigma \) decomposes into \( Q_0 = 2 \) bilinear and \( Q_1 = 1 \) trilinear components,

\[
\begin{align*}
a^1_0(\mathbf{u}, \mathbf{v}) &= -\int_\Omega p \partial_i v_i - \int_\Omega q \partial_i u_i + \lambda \int_\Omega p + \gamma \int_\Omega q - \int_\Omega \theta v_2, & \Theta^1_0(\mu) &= 1, \\
a^2_0(\mathbf{u}, \mathbf{v}) &= \int_\Omega \partial_i u_j \partial_i v_j + \int_\Omega \partial_i \theta \partial_i \chi, & \Theta^2_0(\mu) &= \mu_1, \\
a^1_1(\mathbf{u}, \mathbf{z}, \mathbf{v}) &= \int_\Omega u_i \partial_i z_i v_i + \int_\Omega z_j \partial_i u_i v_j + \int_\Omega z_i \partial_i \theta \chi + \int_\Omega u_i \partial_i \zeta \chi, & \Theta^1_1(\mu) &= 1.
\end{align*}
\]

The right-hand side, which linearly depends on \( \mu_1 \), and the (linear) outputs take the form

\[
\begin{align*}
f^1(\mathbf{v}) &= \int_{\Gamma_h} \chi, & \Theta^1_j(\mu) &= \mu_1, \\
f^1_{(1),0}(\mathbf{v}) &= \int_{\Gamma_h} \chi, & \Theta^1_{j(1)}(\mu) &= 1, \\
f^1_{(2),0}(\mathbf{v}) &= \int_{\Gamma_h} \chi, & \Theta^1_{j(2)}(\mu) &= 1.
\end{align*}
\]

6.2. Reduced basis and supremizer. To achieve faster convergence (cf. [17, 33]), in step 4 of the algorithm in section 5.1, the solution \( \mathbf{u}(\mu) = (u_1, u_2, \chi, p, \lambda) \) is split into its physical components. To ensure solvability of the reduced system, we preliminarily add the constant pressure \((0, 0, 0, 1, 0)\) and Lagrange multiplier \((0, 0, 0, 0, 1)\) to our reduced space. To ensure stability, to each new added pressure mode, we add a supremizer (cf. [26, 27]) \( \sigma(\mathbf{u}(\mu)) \in X \), defined by a null temperature, pressure, and Lagrange multiplier, and

\[
(\sigma(\mathbf{u}(\mu)), \mathbf{v})_X = \int_\Omega p \partial_i v_i \quad \forall \mathbf{v} \in X.
\]

(Note that there are alternative definitions of the supremizer; cf. [26, 27].)

Therefore step 4 is replaced by

4bis. Add the three basis functions \((u_1, u_2, 0, 0, 0), (0, 0, \chi, 0, 0), \) and \((0, 0, 0, p, 0)\)

as well as the supremizer \( \sigma(\mathbf{u}(\mu)) \) to our reduced basis space.

6.3. Sobolev embedding constant. We define a seminorm \( |\cdot|_4 \) as the restriction of the \( L^4 \)-norm to the velocity and temperature fields:

\[
|\mathbf{v}|_4 = \left( \int_\Omega \left( v_1^4 + v_2^4 + \chi^2 \right) \right)^{1/2}.
\]

**Proposition 6.1.** For all \( \mathbf{u}, \mathbf{v}, \) and \( \mathbf{z} \) in \( X \),

\[
|a^1_1(\mathbf{u}, \mathbf{z}, \mathbf{v})| \leq 2 |\mathbf{u}|_4 |\mathbf{z}|_4 |\mathbf{v}|_4.
\]
Proof. We denote that \( u_3 = \theta, v_3 = \chi \), and \( z_3 = \zeta \); we then have
\[
\int_\Omega \sum_{i=1}^{2} \sum_{j=1}^{3} u_i \partial_i z_j v_j \leq \int_\Omega \left( \sum_{j=1}^{3} v_j^2 \right)^{\frac{1}{2}} \left( \sum_{j=1}^{3} \left( \sum_{i=1}^{2} u_i \partial_i z_j \right)^2 \right)^{\frac{1}{4}} \]
\[
\leq \left[ \int_\Omega \left( \sum_{j=1}^{3} v_j^2 \right)^{\frac{1}{2}} \right]^{\frac{3}{4}} \left[ \int_\Omega \left( \sum_{j=1}^{3} \left( \sum_{i=1}^{2} u_i \partial_i z_j \right)^2 \right)^{\frac{1}{2}} \right]^{\frac{1}{4}} \]
\[
\leq \|v\|_4 \left[ \int_\Omega \left( \sum_{i=1}^{2} u_i^2 \right)^{\frac{1}{2}} \right]^{\frac{3}{4}} \left[ \int_\Omega \sum_{j=1}^{3} \left( \partial_i z_j \right)^2 \right]^{\frac{1}{4}} \]
\[
\leq \|v\|_4 \left[ \left[ \int_\Omega \left( \sum_{i=1}^{2} u_i^2 \right)^{\frac{1}{2}} \right] \left[ \int_\Omega \sum_{j=1}^{3} \left( \partial_i z_j \right)^2 \right] \right]^{\frac{1}{4}} \]
\[
\leq \|v\|_4 \left[ \|u_1, u_2\|_{L^4(\Omega)^2} \|(z_1, z_2, z_3)\|_{H^1(\Omega)^3} \right].
\]
From
\[
\sum_{i=1}^{2} \sum_{j=1}^{3} \partial_i (u_j z_i) = \sum_{i=1}^{2} \sum_{j=1}^{3} u_j \partial_i z_i + \sum_{i=1}^{3} \sum_{j=1}^{2} z_i \partial_i u_j,
\]
integration by parts, and \( z_1 = z_2 = 0 \) on \( \partial \Omega \), we have
\[
\int_\Omega \sum_{i=1}^{2} \sum_{j=1}^{3} z_i \partial_i u_j v_j = - \int_\Omega \sum_{i=1}^{2} \sum_{j=1}^{3} u_j z_i \partial_i v_j - \int_\Omega \sum_{i=1}^{2} \sum_{j=1}^{3} u_j z_i \partial_i v_j.
\]
We bound the first term on the right-hand side by applying (6.1) and the second one by
\[
\int_\Omega \sum_{i=1}^{2} \sum_{j=1}^{3} u_j \partial_i z_i v_j \leq \|v\|_4 \left[ \int_\Omega \left( \sum_{i=1}^{2} u_i^2 \right) \sum_{i=1}^{3} \left( \partial_i z_i \right)^2 \right]^{\frac{1}{2}} \]
\[
\leq \|v\|_4 \left[ \int_\Omega \left( \sum_{i=1}^{2} u_i^2 \right)^{\frac{1}{2}} \right]^{\frac{3}{4}} \left[ \int_\Omega \sum_{i=1}^{3} \left( \partial_i z_i \right)^2 \right]^{\frac{1}{4}} = \|v\|_4 \|u\|_4 \|(z_1, z_2)\|_{H^1_{\text{div}}(\Omega)^2}.
\]
\[
\|(u_1, u_2)\|_{L^4(\Omega)^2} \leq \|u\|_4 , \|(z_1, z_2, z_3)\|_{H^1(\Omega)^3} < \|z\|_X, \|(z_1, z_2)\|_{H^1_{\text{div}}(\Omega)^2} < \|z\|_X, \text{ and the definition of } a_1^k \text{ complete the proof.}
\]

From the Sobolev embedding theorem we know that there exists a constant \( \rho_X \) such that
\[
|v|_4 \leq \rho_X \|v\|_X \forall v \in X
\]
and from the equivalence of the norms \( \| \cdot \|_X \) and \( \| \cdot \|_{\overline{P}} \) that there exists a constant \( \rho_\overline{P} \leq \frac{\rho_X}{\|x\|} \) such that
\[
|v|_4 \leq \rho_\overline{P} \|v\|_{\overline{P}} \forall v \in X.
\]
With

\[ \rho_{\mathbf{p}}(\mu) \equiv \frac{\theta_{\mathbf{p}} \rho_{X}(\mu)}{\beta(\mathbf{p})} + 2 \rho_{\mathbf{p}}^2 \quad \text{and} \quad \rho_{X,\mathbf{p}} \equiv 3 \rho_{\mathbf{p}} \rho_{X}, \]

the assumptions (2.8), (2.9), and (2.10) are satisfied.

6.4. Computation of the eigenvalues. The computation of the inf-sup constants \( \beta(\mathbf{p}) \) and of the extreme eigenvalues \( \lambda_{p, \inf}^{\mathbf{p}}, \gamma_{i,q, \inf}^{\mathbf{p}} \), \( p = 1, \ldots, P, q = 1, \ldots, Q_i, i = 0,1 \), heavily depends on the initial guesses. We start by computing \( \beta(\mathbf{p}) \) for all \( \mathbf{p} \) in \( V^K \) from low to high Grashof numbers by inverse iterations. We store all of the modes, and at each \( \mathbf{p} \) we select the initial guess as the stored mode with the smallest Rayleigh quotient.

Then we compute the remaining eigenvalues, also by selecting among the stored modes and by storing the new modes found. In this case, we use a locally optimized block preconditioned gradient method [10]. Figure 2 shows \( \beta(\mathbf{p}) , \lambda_{1, \inf}^{\mathbf{p}}, \gamma_{1,1, \inf}^{\mathbf{p}}, \) and the comparison with the stored modes; some maximizing modes are almost identical for different values of \( \mathbf{p} \).
6.5. Computation of the Sobolev embedding constants. We have developed an algorithm [19] for the computation of the Sobolev embedding constants \( \rho_X \) and \( \rho_\mathcal{P} \) in (6.2) and (6.3).

Let \( u^* \in X, \|u^*\| = 1 \), be the (possibly nonunique) supremizer of

\[
\rho = \sup_{v \in X} \frac{|v|_4}{\|v\|},
\]

where \( \| \cdot \| \) denotes the \( X \) or the natural norms, and let \( z : X \to L^2(\Omega) \),

\[
z(v) = \frac{v_1^2 + v_2^2 + \chi^2}{|v|_4^2}.
\]

Note that \( \|z(v)\|_{L^2(\Omega)} = 1 \) for all \( v \) in \( X \).

For a nonnegative function \( z \in L^2(\Omega) \), we introduce the eigenproblem: Find \( u = (u_1, u_2, \theta, \ldots) \in X \) and \( \lambda \in \mathbb{R} \) such that \( \|u\| = 1 \) and

\[
\int_{\Omega} z(u_1 v_1 + u_2 v_2 + \theta \chi) = \lambda (u, v) \quad \forall v = (v_1, v_2, \chi, \ldots) \in X,
\]

where \((\cdot, \cdot)\) is the \( X \) or natural scalar products. We denote by \( \lambda_{\text{max}}(z) \) and \( u_{\text{max}}(z) \) the largest eigenvalue and associated eigenfunction, respectively. Note that we choose the eigenfunctions always in the same half-space.

We observe that \( \lambda_{\text{max}}(z(u^*)) = \rho^2 \) and \( u_{\text{max}}(z(u^*)) = u^* \), which suggests the following fixed point algorithm:

Let \( u^0 \in X \) be our initial guess; for \( k \geq 0 \), define

\[
u^{k+1} = u_{\text{max}}(z(u^k)),\quad \lambda^{k+1} = \lambda_{\text{max}}(z(u^k)).
\]

A fixed point of this algorithm is not necessarily the supremizer of (6.4), but it is at least a local supremizer. In fact

\[
\lambda^{k+1} - \lambda^k = |u^{k-1}|_4^2 \int_{\Omega} (z(u^{k-1}) - z(u^{k-2}))z(u^{k-1}) + o(||z(u^{k-1}) - z(u^{k-2})||_{L^2(\Omega)}^2),
\]

and, since \( \|z(v)\|_{L^2(\Omega)} = 1 \),

\[
\int_{\Omega} (z(u^{k-1}) - z(u^{k-2}))z(u^{k-1}) \geq 0;
\]

if \( z(u^{k-1}) \) and \( z(u^{k-2}) \) are not parallel (i.e., not equal), this integral is strictly positive.

The most critical point of this algorithm is hence the initial guess. To provide a robust initial guess, we first look for the maximizing mode in the low-dimensional subspace generated by the eigenmodes of \( T_N^\mathcal{P} \) and the already computed supremizers of (6.4). We apply our algorithm in this space, and we select our initial guesses among the basis elements (i.e., among the eigenmodes of \( T_N^\mathcal{P} \) and the supremizers of (6.4)). This process depends on the order that we choose in \( \mathcal{V}^K \); to relax this dependency, when the computations are completed for all \( \mathcal{P} \in \mathcal{V}^K \), we store the modes and restart the computation of all of the embedding constants. The last picture in Figure 2 shows the embedding constants and the comparison with the quotient (6.4) of the stored modes. In Figure 3 note how, for a high Grashof number, \( \rho_\mathcal{P} \) deflates significantly from its upper bound \( \rho_X/\beta(\mu) \).
REDUCED BASIS ERROR BOUND COMPUTATION OF N-S EQUATIONS


7.1. Finite element solver. We implemented a parallel finite element solver that uses the Trilinos [36] library, in particular Epetra as MPI interface, Amesos [31, 30] as dense direct solver, Aztec00 as linear iterative solver with domain decomposition preconditioners from IFPACK [29], and Anasazi as eigenvalue solver, in particular LOBPCG [10].

We exploited the PSC resources [37].

We used Taylor–Hood P1/P2 finite elements for a total of 38000 degrees of freedom. We modified a preconditioner proposed by Elman, David, and Wathen [5]: At a point \( u \in X \), we define \( P^0 \) as \( da(u, \mu) \) where the divergence operator is replaced by the pressure mass matrix on the pressure block diagonal. We then construct a one-level Schwarz preconditioner \( P \) [29, 28] to \( P^0 \) and perform a local LU factorization. We solve the Jacobian system with restarted PGMRES(500) with preconditioner \( P \). The resulting operator \( P \), which depends on \( u \) and \( \mu \), is fast, distributed, and an effective preconditioner for our problem up to \( Gr = 10^7 \).

We are interested in one solution branch; therefore, we use homotopy with respect to the Grashof number when solving the finite element problem. In contrast, in our example, the reduced basis problem does not need homotopy: Our initial guess is a known solution for a nearby parameter (a solution that has generated our reduced space).

In Figures 4 and 5 we present the streamlines and the profiles for an increasing Grashof number. We note the development of two vortices and a boundary layer against the wall. We also show the behavior of the Reynolds number (given by

![Graph showing \( \rho_{H}, \rho_{X}/\beta(\mu), \lambda_{max}, \lambda_{min}, \gamma_{1,1,max}, \gamma_{1,1,min} \) with respect to Grashof number. Notes include similarity for low Grashof numbers and a significant difference at higher numbers, encouraging the use of the mixed formulation of the Brezzi-Rappaz-Raviart theory.]

Fig. 3. Plot of \( \rho_{H}, \rho_{X}/\beta(\mu), \lambda_{max}, \lambda_{min}, \gamma_{1,1,max}, \gamma_{1,1,min} \). Note that \( \rho_{H} \leq \rho_{X}/\beta(\mu) \) and that they are similar for a low Grashof number, but that at a higher Grashof number the difference is almost of one order of magnitude. This encourages one to use the mixed formulation of the Brezzi-Rappaz-Raviart theory.
Fig. 4. Streamlines at Grashof number $10^4$, $10^5$, $10^6$, and $10^7$.

Fig. 5. Profiles at Grashof number $10^4$, $10^5$, $10^6$, and $10^7$. On the left is the vertical central section, on the right the horizontal one.
\[ \text{Fig. 6. Approximative relation between Grashof and Reynolds numbers.} \]

\[ \text{Fig. 7. Outputs } s_{(1)}(\mu) = \frac{1}{|\Gamma_h|} \int_{\Gamma_h} \theta(\mu) \text{ and } s_{(2)}(\mu) = \frac{1}{|\Gamma_s|} \int_{\Gamma_s} u_1(\mu) \text{ with } N = N_{\text{max}} = 2 + 4 \cdot 19. \]

\[ \sqrt{\text{Gr}} \cdot \max(|u_1, u_1|) \text{ with respect to the Grashof number (Figure 6) and our outputs (Figure 7, computed by the reduced basis method).} \]

The offline computation, i.e., the computation of the reduced basis and the inf-sup lower bound ingredients, required about 36 wall-time hours on a 16 CPU cluster.

### 7.2. Reduced basis resolution and error bounds.

Table 1 compares the CPU time needed by an online reduced basis resolution and a finite element resolution. In the latter, we optimize the initial guess by taking the reduced basis solution such that the X-dual norm of the residual is already below $10^{-6}$ and two Newton iterations...
Table 1
Convergence rate and CPU time comparison over 1000 random samples (different for different \( N \)). \( \tau_{N,\Pi}(\mu) = \infty \) means that for some parameters our inf-sup lower bound is negative. The maximal error bounds are taken over the samples for which we can provide existence and unicity. The RBcpu counts the mean CPU time for solving the reduced basis plus the time to compute the error bound (dual norm of the residual and inf-sup lower bounds); the FEcpu counts the mean CPU time for computing the finite element solution with a dual norm of the residual smaller than \( 10^{-10} \) and the reduce basis solution as the initial guess.

<table>
<thead>
<tr>
<th>( N )</th>
<th>max ( \tau_N )</th>
<th>max ( \Delta_N )</th>
<th>max ( \Delta_{(1),N} )</th>
<th>max ( \Delta_{(2),N} )</th>
<th>RBcpu</th>
<th>FEcpu</th>
</tr>
</thead>
<tbody>
<tr>
<td>2+4-10</td>
<td>( \infty )</td>
<td>4.05( \cdot 10^{-4} )</td>
<td>5.60( \cdot 10^{-3} )</td>
<td>4.57( \cdot 10^{-4} )</td>
<td>15+19 ms</td>
<td>2+4-10</td>
</tr>
<tr>
<td>2+4-12</td>
<td>( \infty )</td>
<td>3.60( \cdot 10^{-4} )</td>
<td>4.81( \cdot 10^{-3} )</td>
<td>3.93( \cdot 10^{-4} )</td>
<td>18+35 ms</td>
<td>2+4-12</td>
</tr>
<tr>
<td>2+4-14</td>
<td>2.26</td>
<td>2.94( \cdot 10^{-5} )</td>
<td>6.92( \cdot 10^{-4} )</td>
<td>9.24( \cdot 10^{-5} )</td>
<td>22+66 ms</td>
<td>2+4-14</td>
</tr>
<tr>
<td>2+4-16</td>
<td>0.128</td>
<td>6.32( \cdot 10^{-6} )</td>
<td>3.08( \cdot 10^{-5} )</td>
<td>2.98( \cdot 10^{-5} )</td>
<td>25+108 ms</td>
<td>2+4-16</td>
</tr>
<tr>
<td>2+4-18</td>
<td>0.113</td>
<td>3.16( \cdot 10^{-6} )</td>
<td>1.154( \cdot 10^{-4} )</td>
<td>1.49( \cdot 10^{-5} )</td>
<td>27+157 ms</td>
<td>2+4-18</td>
</tr>
<tr>
<td>2+4-19</td>
<td>0.122</td>
<td>5.12( \cdot 10^{-7} )</td>
<td>3.34( \cdot 10^{-6} )</td>
<td>8.20( \cdot 10^{-6} )</td>
<td>29+188 ms</td>
<td>2+4-19</td>
</tr>
</tbody>
</table>

![Inf–Sup lower bound (log10–scale), \( N = 19 \)](image)

Figure 8. \( \overline{L}^{LB}_{\Pi}(\mu) \), with \( N = N_{\text{max}} = 2+4-19 \) and \( \kappa \) satisfying (4.9).

are usually enough to reduce it to \( 10^{-10} \). Although this optimization is very generous with the finite element code, the wall time is about two orders of magnitude smaller when using the reduced basis method. Moreover, the online reduced basis problem can be solved on a notebook (one CPU and about 1 GB of memory), while the finite element problem has been solved on a cluster using 16 CPUs.

Once \( \tau_{N,\Pi}(\mu) \leq 1 \), our greedy algorithm seeks for a reduction in the error, not in \( \tau_{N,\Pi}(\mu) \) itself; our random set changes at every new search, which explains why in the last two lines of the table \( \tau_{N,\Pi}(\mu) \) increases.

Figure 8 shows that our inf-sup lower bound is everywhere between 0.1 and 1. We also recognize the elements of \( \mathcal{V}^K = \{\overline{\mu}_1, \ldots, \overline{\mu}_{24}\} \) where \( \overline{L}^{LB}_{\Pi}(\mu) = 1 \).

The error bounds for \( N = 2+4-19 \) are small, and, also important, the effectivities are small thanks to our natural norms approach (cf. (3.15) and Figure 9). Note how the error bounds are closely related to the parameters that generate our reduced basis,
while the effectivities are closely related to the inf-sup lower bound (Figure 8). Since our inf-sup lower bound is better “nearby a $\mu$” (cf. (4.1) and (4.2)) the effectivity is better “nearby a $\mu$.”

In Figure 7 we show our outputs $s_{(1)}$ and $s_{(2)}$ with respect to the Grashof number, computed with our reduced basis on 1000 random points; the error bound is always smaller than $5 \cdot 10^{-5}$. The effectivities for the first output (Figure 10) are less sharp than for the field variable; anyway, when the error bound is small, a “bad” effectivity is acceptable, and when the error bound is the largest, the effectivity is good.
8. Conclusions. We have modified the Brezzi–Rappaz–Raviart theory and defined the natural norm in the quadratic case; the latter enables a feasible and locally sharp inf-sup lower bound construction. These three components provide a fast and feasible error bound computation, which is essential in both the online and the offline stages of the reduced basis method. Our approach has proven good effectivities in the field variable error bounds and empirical good effectivities in the output error bounds.

One essential point is that the error bounds—with respect to the truth finite element solution—are exact; hence, the computation of the inf-sup lower bound components have to be exact, too. We have developed an algorithm to remedy a robust Sobolev embedding constant computation.

We are hence able, first, to prepare all of the inf-sup lower bound components offline with a clear methodology and in a reasonable CPU wall time and, second, to solve online quadratic partial differential equations—natural convection with a Grashof number up to $10^7$—and provide fast and reliable error bounds, with gains in term of online resources—a notebook against a 16 processor cluster—and in terms of CPU wall time—two orders of magnitude.

Future work will be to show that even in the multiparameter case our approach is feasible; i.e., the offline CPU time is acceptable and the online performance is not affected (preliminary tests with Grashof and Prandl numbers as parameters are promising), including in the presence of geometrical changes (where, e.g., the aspect ratio of the cavity is a further parameter).

We have presumed that the (parameter-dependent) solution lays on an isolated branch. The natural norm approach should also allow us to consider bifurcations since with respect to the “local” norm the inf-sup constant is of order one. Nevertheless, the natural norm degenerates near the bifurcation because of a singular mode. This mode probably has to be treated by deflation, and the eigenvalue solver should take into account the norm degeneration.

Appendix. Dual approximation. In addition to our “truth” primal problem, we may also consider a “truth” dual (or adjoint) problem [23, 14, 3, 21, 33, 32] associated with our particular output functional to improve the convergence of the output.

In this section we consider the case of the quadratic, parameter-independent output (the extension to the linear or the affine-dependent cases is straightforward)

\[
\ell(v) = \ell_0(v) - \ell_1(v,v),
\]

where \(\ell_0\) is continuous and linear and \(\ell_1\) continuous, symmetric, and bilinear.

For a given approximation \(u_N(\mu)\) in \(W_N\) to \(u(\mu)\), the dual problem reads: Find \(\psi_N^N\) in \(X\) such that

\[
(A.1) \quad da(u_N(\mu); \mu)(v, \psi_N^N(\mu)) = -\ell_0(v) - 2\ell_1(u_N(\mu), v) \quad \forall v \in X.
\]

We apply our reduced basis method to solve the dual problem \((A.1)\) and solve for the reduced basis solution \(\psi_N^{du}(\mu)\) in \(W_{du}^N\) such that

\[
\text{da}(u_N(\mu); \mu)(v, \psi_N^{du}(\mu)) = -\ell_0(v) - 2\ell_1(u_N(\mu), v) \quad \forall v \in W_{du}^N,
\]

where \(W_{du}^N\) (in general different from \(W_N\)) is a reduced basis space of dimension \(N\) (for simplicity in this presentation, we consider the same dimension for the reduced
primal and dual spaces). We denote the primal and dual residuals as
\[ r(v; \mu) = a(u(\mu), v; \mu) - a(u_N(\mu), v; \mu) = f(v) - a(u_N(\mu), v; \mu) \quad \forall v \in X, \]
\[ r^{N,du}(v; \mu) = da(u_N(\mu); \mu)(v, \psi^N(\mu)) - da(u_N(\mu); \mu)(v, \psi^N_N(\mu)) = -\ell_0(v) - 2\ell_1(u_N(\mu), v) - da(u_N(\mu); \mu)(v, \psi^N_N(\mu)) \quad \forall v \in X, \]
respectively, and the primal error as \( e(\mu) = u(\mu) - u_N(\mu) \). We define the dual corrected output as
\[ s^{N,du}_N(\mu) = \ell(u_N(\mu)) - r(\psi^N_N(\mu); \mu). \]

The following lemma holds.

**Lemma A.1.**

\[ s(\mu) - s^{N,du}_N(\mu) \]
\[ = \ell_1(e(\mu), e(\mu)) - r^{N,du}(e(\mu); \mu) + \frac{1}{2} \left[ da(u(\mu); \mu) - da(u_N(\mu); \mu) \right] (e(\mu), \psi^N_N(\mu)). \]

**Proof.**
\[ s(\mu) - s^{N,du}_N(\mu) = \ell_0(u(\mu)) - \ell_0(u_N(\mu)) + \ell_1(u(\mu) - u_N(\mu), u(\mu) - u_N(\mu)) \]
\[ + 2\ell_1(u(\mu) - u_N(\mu), u_N(\mu)) + r(\psi^N(\mu); \mu) \]
\[ = \ell_0(e(\mu)) + \ell_1(e(\mu), e(\mu)) + 2\ell_1(u_N(\mu), e(\mu)) \]
\[ + a(u(\mu), \psi^N_N(\mu); \mu) \equiv -\ell_1(e(\mu), e(\mu)) - r^{N,du}(e(\mu); \mu) - da(u_N(\mu); \mu) (e(\mu), \psi^N_N(\mu)) \]
\[ + \frac{1}{2} \left[ da(u(\mu); \mu) + da(u_N(\mu); \mu) \right] (e(\mu), \psi^N_N(\mu)). \]

We are now able to compute an upper bound for the dual corrected output. Let

\[ \| \ell_1 \|_\pi \equiv \sup_{v \in X} \frac{\| \ell_1(v, v) \|}{\| v \|_\pi^2}, \]
\[ \| r^{N,du}(\cdot; \mu) \|_\pi \equiv \sup_{v \in X} \frac{\| r^{N,du}(v; \mu) \|}{\| v \|_\pi}, \]

and

\[ \Delta^s_{N,\pi}(\mu) \equiv \| r^{N,du}(\cdot; \mu) \|_\pi^2 \Delta_N(\pi(\mu) + \frac{1}{2} \beta_{\pi}(\mu) \psi^N_N(\mu) + \ell_1 \|_\pi^2 \Delta_N(\pi(\mu))^2. \]

Then, from (2.8), (3.14), (A.2), and (A.3), the following output error bound holds.

**Theorem A.2.**

\[ |s(\mu) - s^{N,du}_N(\mu)| \leq \Delta^s_{N,\pi}(\mu). \]

As we expect from the dual approach, the error bound is quadratic with respect to the dual norm of the residuals \( r \) and \( r^{N,du} \). The main advantage of (A.4) is that there is no dependence on \( \beta(\pi) \), which is the square root of the worse mode of \( \pi_N \).
however, this mode may be hidden behind other terms like the natural dual norm of the dual residual or \( \mathcal{r}(\mu) \) (both explicitly) or in \( \Delta_N(\mu) \).

Note that, to simplify the offline work, the term \( \| \mathcal{r}(\mu) \| \) may also be replaced by using

\[
\| \mathcal{r}(\mu) \| \leq \frac{1}{\beta(\mu)} \| \mathcal{r}(\mu) \|_{X'}
\]

at the detriment of the independency on \( \beta(\mu) \).

Acknowledgments. I thank Professor Anthony T. Patera for his valuable comments and many contributions. I am grateful to Dr. Karen Veroy–Grepl and Dr. Gianluigi Rozza for many discussions on the subject.

REFERENCES


A. T. Patera, private communication.


http://www.psc.edu.